

Abstract

In a computer data processing system, method and apparatus for clustering data in a database. A database having a number of data records having both discrete and continuous attributes is stored on one or more storage media which may be connected by a network. The records in the database are scanned so that data records which have the same discrete attribute configuration can be tabulated. A first set of configurations is determined wherein the number of data records of each configuration of said first set of configurations exceeds a threshold number of data records. Data records that do not belong to one of the first set of configurations are added to or tabulated with a configuration within said first set of configurations to produce a subset of records from the database belonging to configurations in the first set of configurations. The data in these configurations are then clustered based on the continuous data attributes of records contained within that first set of configurations to produce a clustering model.